J Indian Inst Sci., Mar –Apr. 2000, **80**, 143–177

IISc THESES ABSTRACTS

Thesis Abstract (Ph. D.)

Structural motifs in designed synthetic peptides: NMR characterisation by S. Raghothama
Research supervisors: Profs P. Balaram and C. L. Khetrapal
Department: Molecular Biophysics Unit

1. Introduction

The design and characterisation of synthetic peptide modules having well-defined secondary structure motifs play an important role in further assembly of these modules leading to tertiary structures. The designed peptides mimic structural features found in proteins. NMR spectroscopy is often the method of choice in such conformational analysis. NMR parameters have diagnostic features which can distinguish among various secondary structures. Information on tertiary interactions can also be obtained by long-range NOE interactions. These NMR parameters provide constraints to refine structures by molecular dynamics simulations.

This work describes the design and characterisation of supersecondary structure motifs in peptides ranging from 8–25 residues, using NMR techniques. Specific features such as helix unfolding, β-turn interconversions, use of diagnostic sidechain-backbone NOEs in determining the α-aminoisobutyric acid (Aib) residue conformation are discussed.

2. Results and discussion

Designing a β-hairpin remained a challenge till recently[1–3] though they formed one of the simplest supersecondary structures. An octapeptide Boc-Leu-Val-Val-^{D}Pro-Gly-Leu-Val-Val-OMe was successfully designed and characterised by NMR to form a β-hairpin. The ^{D}Pro-Gly segment nucleated the crucial type II′ β-turn, while the choice of Leu and Val facilitated the extended strand conformations. NMR studies in $CDCl_3$ and C_6D_6 solvent clearly established a β-hairpin structure, while in DMSO two conformations corresponding to the *cis* and *trans* geometry about the X-Pro bond, were established, of which *trans* conformation was shown to be a β-hairpin. Molecular dynamics simulations, using NMR-derived constraints, led to a family of closely related hairpin conformations (Fig. 1). The studies were compared with an analog peptide containing ^{L}pro-Gly segment. Further design and characterisation was extended to a four-stranded 25-residue peptide having three ^{D}pro-Gly segments.

Helix-loop-helix (α, α) motif is another important supersecondary structure. A modular approach known as a Mecano-set approach[4, 5] is being developed in our laboratory that envisages stepwise assembly of conformationally rigid helices into $\alpha\alpha$ motifs by means of intervening linker segment. Studies are carried out on a series of 15/16 residue peptides made up of two helical hepta-peptide segments interlinked by an Acp (ε-aminocaproic acid) residue. These heptapeptides contain α-aminoisobutyric acid (Aib) residue, which are well-known stabilizers of helices. The solubility of these peptides in relatively inert solvents like chloroform and

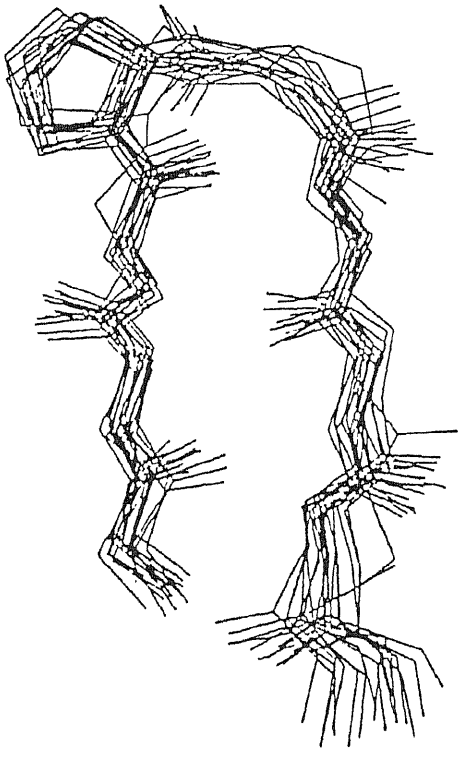

FIG. 1. Superposition of 20 structures collected from restrained molecular dynamics run on peptide Boc-Leu-Val-Val-^{D}Pro-Gly-Leu-Val-Val-OMe.

methanol permits folding processes to be studied in the absence of a strong hydrophobic driving force.

In all peptides, helical segments were clearly identified by a continuous stretch of sequential d_{NN} NOEs. The observation of long-range NOEs which should in principle establish the compact antiparallel arrangement was restricted to the region around the central Acp hinge. This indicated a bent arrangement of the helices, precluding small interhelical angles. The introduction of ^{D}Pro residue into the linking segment along with Acp residue was further considered to stereochemically drive the structure in the peptide Boc-Val-Ala-Leu-Aib-Val-Ala-Leu-Acp-^{D}Pro-Leu-Aib-Val-Ala-Leu-Aib-Val-OMe. NMR analysis resulted, as in previous cases, interhelical NOEs, centered around the hinge region. A model built using these limited NMR constraints suggested an orthogonal arrangement of the helices, consistent with the experimental data (Fig. 2).

The achiral, stereochemically constrained Aib residue has been extensively used in the design of peptides adapting well-defined conformations.[6–8] Studies were carried out to understand the usefulness of sidechain-backbone NOE which can serve as a diagnostic tool to determine

the local conformation of an Aib residue in peptides. Though Aib has a pronounced energy minimum in the helical region of ϕ, ψ space, there are indeed other minima which are slightly higher in energy, corresponding to semi- and fully extended conformations, examples of which exist in the literature.[6] Analysis of the dependence of intra- and inter-residue distances between backbone NH and Aib $C^{\beta}H_3$ protons suggests that $d_{\beta N}(i, i+1)$ [Aib $C^{\beta}H_3 \leftrightarrow$NH] NOEs may be useful diagnostics in determining the conformation of this residue. Further, the usefulness of these NOEs in detecting conformational averaging which are important in solution studies is also demonstrated using model peptides examples.

Helix unfolding in an acyclic octapeptide, a conformational transition induced by changes in solvent composition was studied using NMR spectroscopy. Incorporation of nonprotein amino acids, in particular α, α-dialkylated residues, may be used to impart conformational rigidity to polypeptide backbone.[5] One such particular residue is the 1-aminocyclooctane-1-carboxylic acid (Ac_8c) which resembles Aib in the conformational characteristics imparting helical conformations in oligopeptides.[9] Incorporation of this residue at position 2 and 7 in an octapeptide Boc-Leu-Ac_8c-Val-Gly-Gly-Leu-Ac_8c-Val-OMe constrains the local residue con-

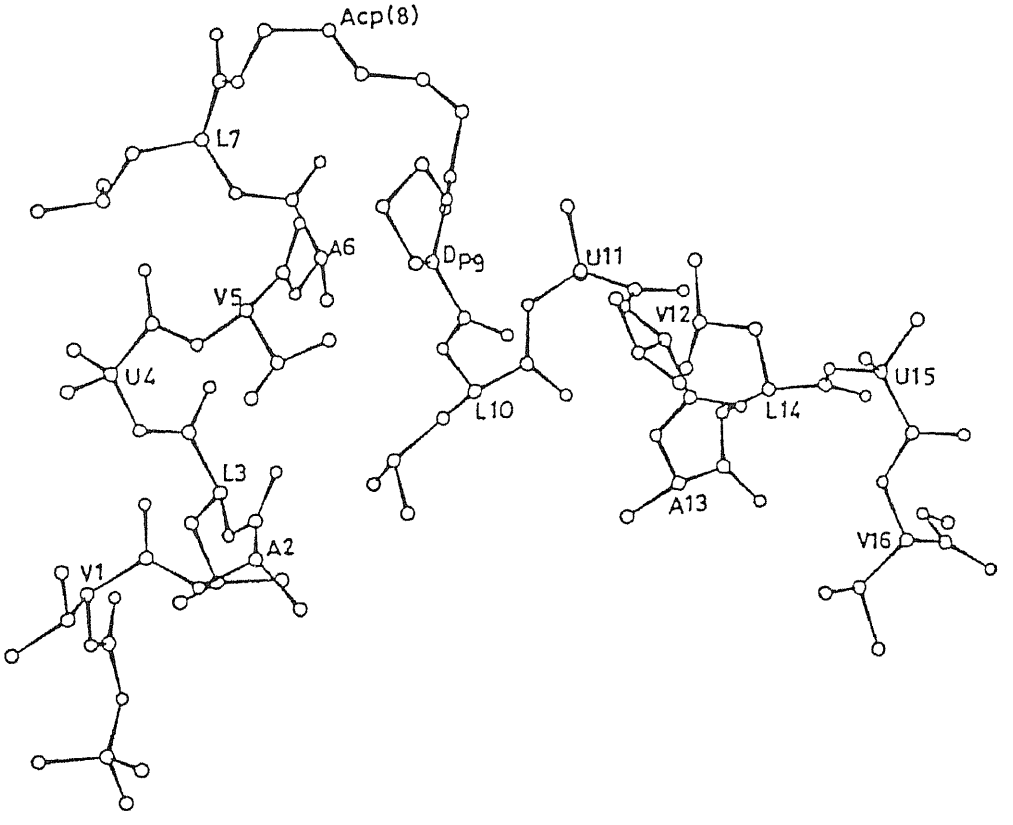

FIG. 2. Restrained energy minimized model of a 16-residue peptide.

formation to the right or left-handed helical region of ϕ, ψ space. The central Gly-Gly segment was chosen to provide an element of structural flexibility. The juxtaposition of residues with contrasting stereochemical properties permits the setting up of a model situation where a conformational transition is realized by modulating solvent conditions. The NMR studies establish a conversion from a 3_{10}-helical conformation in a poorly hydrogen-bonding solvent $CDCl_3$ to a multiple turn conformation in a strongly hydrogen-bonding solvent DMSO. Studies in $CDCl_3$/DMSO mixtures show clear evidence for a solvent-dependent conformational transition. Amide NH chemical shift and temperature coefficients determined over a range of solvent composition allows a detailed structural analysis of the unfolding process. The present example suggests that the use of conformationally constrained and flexible residues in conjunction can lead to the design of peptide systems that are finely poised to undergo conformational transitions.

Peptide models for β-turns have attracted current interest in view of their importance in understanding fundamental features of peptide conformation.[10–12] Studies on conformational interconversions in peptide β-turns were carried out. It is established that the introduction of a chiral center adjacent to a β-turn composed of two achiral residues (Aib-Gly) in the peptide Boc-Aib-Gly-Leu-OMe permits spectroscopic detection of conformational interconversion between enantiomeric turn structures. These interconversions are characterised by low barriers as evidenced by dynamic averaging at moderate temperatures. The achiral Aib-Gly segment favours either type I/III or the enantiomeric type I′/III′ conformations. These mirror image β-turn conformations are isoenergetic in achiral sequences. The induction of a chiral residue, L-Leu, at position 3 makes the two types of conformations diastereomeric and thus energetically non-equivalent and provides a handle of spectroscopic distinction. Since the chiral centre is placed outside the central positions of the β-turn, it may be expected that both conformational types may be present in appreciable populations. In crystals, two conformational diastereomers are characterised in the asymmetric unit. The presence of diastereomeric conformations in equal proportions in the solid state is also confirmed by solid-state ^{13}C NMR studies. In solution, evidence for conformational interconversion is presented using variable temperature ^{1}H-NMR.

The studies thus described illustrate the usefulness of NMR spectroscopy in various aspects of conformational analysis, such as characterisation of supersecondary structure motifs in designed peptides, probing conformational dynamics and in establishing local conformations at nonprotein amino acid residues in synthetic peptides. The conformational characterisation of *de novo*-designed peptides is an essential element in further developing strategies for the use of specific nonstandard amino acids in synthetic protein design.

References

1. HAQUE, T. S., LITTLE, J. C AND GELLMAN, S. H. — *J. Am. Chem. Soc*, 1994, **116**, 4105–4106.
2. AWASTHI, S. K., RAGHOTHAMA, S. AND BALARAM, P — *Biochem Biophys. Res. Commun.*, 1995, **216**, 375–381.
3. KARLE, I. L., AWASTHI, S. K. AND BALARAM, P. — *Proc. Natn Acad. Sci. USA*, 1996, **93**, 8189–8193.

4.	BALARAM, P.	*Pure Appl. Chem*, 1992, **64**, 1061–1066.
5.	BALARAM, P	*Curr. Opin Struct. Biol*, 1992, **2**, 845–851
6.	KARLE, I. L. AND BALARAM, P.	*Biochemistry*, 1990, **29**, 6747–6756
7.	MARSHALL, G. R *et al.*	*Proc. Natn. Acad. Sci. USA*, 1990, **87**, 487–491.
8	TONIOLO, C. AND BEBEDETTI, E.	*Biopolymers*, 1991, **22**, 205–215.
9.	MORETTO, V *et al.*	*J. Peptide Sci*, 1996, **2**, 14–27.
10	WILMOT, C. M AND THORNTON, J. M.	*J. Mol Biol.*, 1988, **203**, 221–232.
11	IMPERIALI, B, FISHER, S. L, MOATS, R. A. AND PRINS, T J.	*J. Am. Chem. Soc*, 1992, **114**, 3182–3188.
12.	CHALMERS, D. K. AND MARSHALL, G. R.	*J Am. Chem. Soc*, 1995, **117**, 5927–5937.

Thesis Abstract (Ph.D.)

Satellite imagery to assess species diversity within a landscape context—Studies in the Western Ghats of India by Harini Nagendra
Research supervisor: Prof. Madhav Gadgil
Department: Centre for Ecological Sciences

1. Introduction

The International Convention on Biological Diversity commits all parties, including India, to inventory, monitor and conserve their biodiversity resources. This is an enormous task, even for a part of the country such as the Western Ghats – a hill chain running parallel to the west coast for over 1600 km, considered as one of the world's biodiversity 'hot-spots' (lat. 8°–21°N, long. 73°–77°E). Remote sensing has tremendous potential for this purpose, as it can provide information about the structure and possibly composition of vegetation over large areas at a glance.

However, given India's heterogeneous and species-rich landscapes, direct mapping of stands of individual plants is not possible. Remote sensing can instead be used to map the distribution of ecosystem types, or landscape element types (LSE types). These maps can then be correlated with species distributions within LSE types, to derive information about diversity at the species level of the Western Ghats. It is however necessary to assess the extent to which a purely remote-sensing-based classification of LSE types can provide information about species distributions. This work investigates this approach, using the Indian Remote Sensing Satellite (IRS) 1B imagery.

2. Methods and results

I approach this problem through investigations at three different spatial scales. First, distribution parameters (mean, standard deviation, skew and kurtosis) of the Normalized Difference Vegetation Index, which is believed to be correlated with vegetation biomass and vigour, are used to map the Western Ghats and the west coast of India, an area of 170,000 sq. km, at a broad ($1:10^6$) scale into different types of landscapes.[1]

An unsupervised classification is carried out, followed by the merger of smaller patches with the most prevalent landscape type in the vicinity, to assign the region to a remaining 205 patches belonging to 11 types of landscapes. The distribution of these 11 types is then compared with topography, rainfall, temperature, population, agriculture and vegetation data for interpretation. It is suggested that a sample landscape be mapped and species distributions studied in each of these patches. Such data can then be extrapolated to obtain information about species diversity in the Western Ghats.

At the second spatial scale, detailed landscape mapping at a scale of 1:25,000 is carried out in 12 landscapes distributed across the Ghats, 10–50 sq. km in area, using supervised classification, with initial field input, and unsupervised classification, without such input.[2] These 12 landscapes belong to five of the 11 landscape types which were mapped across the Ghats. A total of 24 LSE types are encountered in these 12 landscapes.

The accuracy of unsupervised classification is found to be much lower than that of supervised classification. Landscape and LSE-type characteristics like landscape diversity, patch size, patch shape and distance to the nearest neighbour of the same type, calculated using the supervised classification differ significantly from those calculated using unsupervised classification. Unsupervised classification at this scale, therefore, does not provide accurate information, either for landscape mapping or for deriving information about landscape characteristics.

However, within-landscape type variation in landscape characteristics is less than between type variation, for all landscape characteristics considered. This indicates that the NDVI-based unsupervised classification carried out at a broader scale is useful in differentiating landscapes of different types, with different landscape characteristics.

The 24 LSE types encountered in these 12 landscapes are clustered based on their patch areas, shape and nearest neighbour distance, to understand the relationship between them. No distinct grouping of LSE types can be discerned, probably because inter-landscape variation in LSE-type characteristics is very high. This suggests that the influence of the landscape in determining patch characteristics is more than that of the LSE-type.

At the third spatial scale, a 30 sq. km landscape in the Ghats (lat. 14°16′–14°19′N, long. 74°52′–74°54′ E) is mapped into seven LSE types, by supervised as well as unsupervised classification.[3] In this, all Angiosperms (excluding grasses) distributed in these seven LSE types are surveyed in the field using 246 quadrats of 10×10 m, in order to assess whether these types could be distinguished on the basis of their species composition.

LSE types as identified in the field and using supervised classification do harbour significantly distinctive sets of flowering plants, whereas unsupervised classification does not permit classification of LSE types with a high enough degree of accuracy to achieve this. LSE types coupled to satellite imagery are therefore a useful device for organizing a program of assessing and monitoring species diversity.

An important component of monitoring is to assess the efficacy of conservation efforts. A methodology is suggested for this purpose and applied to this landscape. First, based on interviews with local informants, a Landscape Transformation Matrix is prepared, describing the projected probabilities of transformation over the next five years. LSE types are then assigned

conservation values as the sum of values (evergreenness, endemicity to the Ghats, medicinal nature or being a wild relative of cultivated plants) of the species which they harbour.

Finally, for each transformation from one LSE type to another, the desirability of transformation is calculated as the product of the likelihood of its occurrence and the gain/loss in value which will result. These desirabilities of transformation could serve as useful inputs to include biodiversity considerations into developmental planning at the local level.

3. Discussion

The methodology proposed in this work, of broad-scale landscape mapping coupled with point sampling of LSE type and species distribution, is therefore a useful one for assessing and monitoring species diversity in the Western Ghats. This is possibly the first exercise in which methodology for an exercise of biodiversity assessment at different spatial scales has been formulated and tested, using a combination of satellite imagery and ground-based species sampling, and linkages between information at these various scales established.

Based on these results, a proposal for biodiversity assessment, monitoring and conservation in the Western Ghats of India is suggested. There are of course questions which still need to be answered, in order to repeat this exercise at a larger, Ghats-wide scale. These are also discussed.

References

1. NAGENDRA, H AND GADGIL, M. Linking regional and landscape scales for assessing biodiversity: A case study from Western Ghats, *Curr. Sci.*, 1998, **75**, 264–271.

2. NAGENDRA, H. AND GADGIL, M Biodiversity assessment at multiple scales Linking remotely sensed data with field information, *Proc Natn Acad Sci. USA*, 1999, **96**, 9154–9158

3. NAGENDRA, H AND GADGIL, M. Satellite imagery as a tool for monitoring species diversity. An assessment, *J. Appl. Ecol.*, *1999*, **36**, 388–397.

Thesis Abstract (Ph.D.)

Organization of work in the primitively eusocial wasp *Ropalida marginata* by Dhrubajyoti Naug
Research supervisor: Prof. Raghavendra Gadagkar
Department: Centre for Ecological Sciences

1. Introduction

Division of labor is fundamental to colony organization in social insects and the central problem of division of labor revolves around flexibility. Small insect societies like those of the primitively eusocial wasp *Ropalidia marginata* are subject to a high degree of stochastic environmental fluctuations. The adults of these wasps lacking morphological caste differentiation exhibit considerable flexibility in their social roles and therefore these societies are es-

pecially attractive model systems to study organization of work. Moreover, since individual-level selection is considered to be more predominant than colony-level selection in primitively eusocial societies like that of *R. marginata*, it is interesting to study the nature of work organization, which is generally assumed to be a product of colony-level selection.

2. Methods

The relation of age with division of labor was therefore assessed in *R. marginata*. The performance of four functionally significant tasks was analyzed. It was found that age has a definite correlation with division of labor since wasps performed tasks in a distinct sequence in their life with successive tasks being initiated at significantly older ages. Age of a wasp was measured in absolute terms and also relative to other individuals in the colony (age rank). Probability of performance of a given task relative to other tasks (PTP) and absolute rates at which tasks were performed per unit time (FTP) both showed clear age-dependent patterns, confirming the association of age with division of labor. Variance explained for both PTP and FTP was significantly higher with relative age than with absolute age suggesting the former rather than the latter is more important in determining the task of an individual. Inter-individual interactions were found to be a potential mechanism through which wasps can determine their relative age.

3. Results

The ability of inter-individual interactions to regulate age polyethism in social insects was further evaluated by developing a computer simulation model of division of labor. The verbal activator–inhibitor model of Huang and Robinson[1] was formalized and elaborated, using empirically derived parameter values from the study described above. The ages and proportions of individuals performing different tasks in colonies with various age distributions and demand levels were computed. The model generated a clear age polyethism which was flexible enough to provide precocious foragers in colonies with only young individuals and overaged nurses in colonies consisting of only old individuals. The model also showed how workers can respond to changing demand levels by appropriately adjusting the ages and proportions of individuals engaged in various tasks. These results inspire confidence in the idea that adaptive age demography can efficiently regulate division of labor.

The roles of absolute and relative age in division of labor was further discerned by using colonies of *R. marginata* consisting of only young individuals (cohort colonies). The cohort colonies had precocious foragers which exhibited a significantly higher PTP and FTP for estranidal tasks compared to individuals of similar age in normal colonies. This shows that behavioural development can be faster in the absence of older individuals and that wasps can work independent of their absolute age. In general, the division of labor among the wasps in the cohort colonies was similar to that seen among the complete set of individuals in normal colonies but significantly different from that seen among individuals of similar age in normal colonies. The results also showed that relative age may be more important in regulating PTP and absolute age in regulating FTP. A cohesive picture of work organization with two measures each of task performance and age show how the constraint posed by absolute age can be partially overcome by relative age. The results confirm the role of relative age in the division of labor and also many of the predictions of the model described above.

4. Conclusions

This study perhaps represents the strongest demonstration so far of the relation of age with division of labor for any primitively eusocial species. The strong relation of age with division of labor seen in this species suggests that age polyethism can evolve even before workers have lost their reproductive options. The flexibility demonstrated by an age polyethism dependent on relative age shows that age can be a sufficiently flexible rule for division of labor.

References

1 HUANG, Z. AND ROBINSON, G. E. *Proc. Natn Acad. Sci USA*, 1992, **89**, 11726–11729

Thesis Abstract (Ph.D.)

Investigations of nanotubes and other carbon forms by Rahul Sen
Research supervisor: Prof. C. N. R. Rao
Department: Materials Research Centre

1. Introduction

The discovery of nanotubes on the negative electrode during the arc-evaporation of graphite rods has given an added impetus to the study of various forms of carbon.[1] Multi-walled carbon nanotubes (MWNT) are concentric graphitic cylinders closed at either end due to the presence of five-membered rings. Single-walled nanotubes (SWNT), consisting of only a single cylinder of graphite, have also been prepared.[2, 3] Since the discovery of carbon nanotubes several ways of preparing the nanotubes have been explored. Besides the conventional arc-evaporation technique, decomposition of hydrocarbons under inert conditions over metal catalysts enables the formation of nanotubes.[4, 5] The presence of metal particles is essential for the formation of nanotubes by the pyrolysis of hydrocarbons and the diameter of the nanotubes appears to be determined by the size of the metal particles. Ever since the discovery of carbon nanotubes, there has been much interest in preparing nanotubes of other materials, in particular those of boron-nitrogen (B-N), boron-carbon-nitrogen (B-C-N) and boron-carbon (B-C).[6, 7] It was of interest to explore whether carbon nanotubes containing nitrogen (C-N nanotubes) could be prepared. In the present work, formation of carbon nanotubes by the pyrolysis of precursor molecules containing metal and carbon has been investigated in detail, to find new ways of making carbon nanotubes and also to establish the role of metal particles.[8, 9] Besides MWNT, aligned nanotube bundles, which are of vital technological importance, have been prepared from ferrocene pyrolysis.[10] B-C-N, C-N and B-N nanotubes have been prepared by the pyrolysis of appropriate precursor molecules over Co nanoparticles in an Ar atmosphere and the nanotubes so obtained were characterized by various spectroscopic methods.[11, 12] Good yields of SWNT have been obtained by the arc-discharge method as well as pyrolysis of $Fe(CO)_5$ with acetylene under dilute conditions. Carbon anion-like structures containing transition metal particles obtained by arc-discharge as well as precursor pyrolysis have been investigated by electron microscopy and magnetic measurements. Diamond-graphite hybrid structures, containing mixtures of sp^3 and sp^2 carbons have been studied by molecular mechanics and their band gaps estimated by extended Huckel theory (EHT).[13] Hydrogenation of small aromatic

hydrocarbons has been investigated at the semi-empirical AM1/RHF level, because of its relevance to diamond-graphite transformation.[14] Structures and images of derivatives of carbon nanotubes and novel fullerene- and tube-like structures with 5- and 7-membered rings have also been studied by molecular mechanics.[15]

2. Experimental

MWNT were prepared by setting up an arc between graphite electrodes in an atmosphere of 650–700 torr helium in a water-cooled stainless steel chamber. For SWNT synthesis, the cathode comprises a hollow graphite rod filled with a mixture of Y_2O_3, Ni and graphite powder. Pyrolysis experiments were carried out in quartz tube flow reactor located in a horizontal tube furnace. The flow rate of the gases was monitored with mass flow controllers. Pyrolysis of hydrocarbons such as methane and benzene was carried out over Ni or $Ni/AlPO_4$ catalysts, and of precursor molecules such as ferrocene, cobaltocene, nickelocene and $Fe(CO)_5$ under different conditions to yield various types of nanotubes. C-N nanotubes were prepared by the pyrolysis of pyridine, methyl-pyrimidine and triazine over Co nanoparticles under Ar atmosphere. Trimethylamine-borane and ammonia-borane complexes were pyrolysed over Co nanoparticles to obtain B-C-N and B-N nanotubes. All the nanotubes were characterized by TEM and SEM. The C-N and B-C-N nanotubes were also characterized by electron energy loss spectroscopy (EELS) which is done with a spectrometer attached to a transmission electron microscope. This gives compositional analysis on individual nanotubes. XPS, Raman and scanning tunneling spectroscopy were also used to characterize the nanotubes.

3. Results and discussion

Pyrolysis of CH_4 and C_6H_6 in a hydrogen atmosphere in the absence of any catalyst yields mono-dispersed spherical carbon particles. However, if the pyrolysis was carried out under reductive conditions over transition metals (like Ni) or supported transition metal catalysts (like $Ni/AlPO_4$), carbon nanotubes are obtained. This study establishes the catalytic role of transition metal particles for the growth of carbon nanotubes prepared by the pyrolysis of hydrocarbons. The pyrolysis of metallocenes such as ferrocene, cobaltocene and nickelocene in Ar or in Ar-H_2 mixture at 900°C yields carbon nanotubes, some of which are completely filled with metal. Furthermore, pyrolysis of benzene in the presence of small proportion of a metallocene yields large quantities of carbon nanotubes. These results demonstrate that the metal clusters generated from metallocenes act as nucleating centers for the formation and growth of the nanotubes. By varying the conditions of the pyrolysis of ferrocene and carrying out the process in the presence of acetylene, which acts as an additional carbon source, large quantities of aligned-nanotube bundles have been obtained. The ready formation of aligned-nanotube bundles is of vital technological interest today. Pyrolysis of $Fe(CO)_5$ in the presence of benzene or acetylene also gives high yields of carbon nanotubes.

Carbon–nitrogen (C-N) nanotubes have been prepared by the pyrolysis of pyridine, methyl pyrimidine and triazine over Co nanoparticles under Ar atmosphere. The presence of nitrogen in these nanotubes was confirmed by EELS carried out with TEM. In Fig. 1, we show the EEL spectra of two different nanotubes obtained by pyridine pyrolysis. The spectra show characteristic edges at 284 and 400 eV corresponding to K-shell ionization of carbon and nitrogen, respectively. The spectra shown give a stoichiometry of $C_{33}N$ (a) and of $C_{11}N$ for the nanotubes

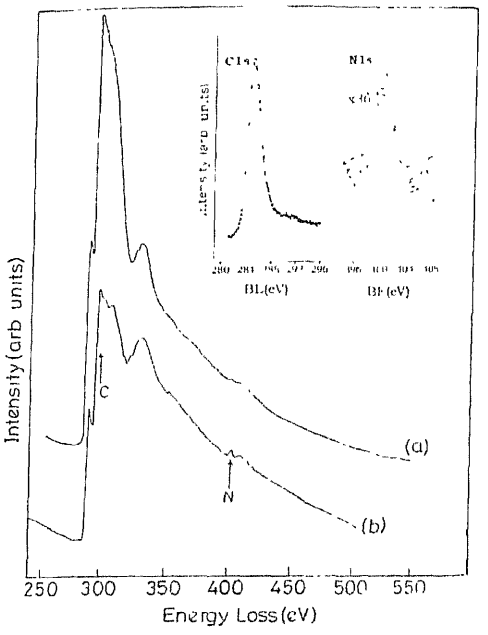

FIG 1 EELS spectra of C-N nanotubes: The spectrum shown in (a) gives a stoichiometry of $C_{33}N$ and that of (b) a stoichiometry of $C_{11}N$ The inset shows the C1s and N 1s signals obtained from XPS of C-N nanotubes.

(b). In the inset of Fig. 1, we show the core-level XPS of the nanotubes obtained by pyridine pyrolysis. The stoichiometry of the C-N nanotubes from XPS band intensities comes out to be $C_{38}N$, close to the composition from EELS in Fig. 1(a). Raman and scanning tunneling spectroscopy show these nanotubes to be distinctly different from carbon nanotubes.

A comparative study of nanotubes of B-C-N and B-N prepared by the pyrolysis of 1:1 addition compounds of BH_3 with $(CH_3)_3N$ and NH_3 has been carried out. B-C-N nanotubes have been characterized by XPS and EELS to determine their composition. There is a significant compositional variation in a given batch of B-C-N nanotubes. The near absence of B-N nanotubes on pyrolysing the appropriate precursor compound and other observations made in the present study indicate the crucial role of carbon in the initial nucleation and growth of nanotubes.

References

1. IIJIMA, S. *Nature*, 1991, **354**, 56-58.
2. IIJIMA, S. AND ICHIHASHI, T. *Nature*, 1993, **363**, 603–605.
3. BETHUNE, D. S. *et al.* *Nature*, 1993, **363**, 605–607.
4. JOSE-YACAMAN, M., MIKI-YOSHIDA, M., RENDON, L. AND SANTIESTEBAN, T. G. *Appl. Phys. Lett.*, 1993, **62**, 202–204.

5.	IVANOV, V. *et al.*	*Chem. Phys. Lett*, 1994, **223**, 329–333
6.	TERRONES, M. *et al.*	*Chem Phys. Lett.*, 1996, **259**, 568–573.
7.	ZHANG, Y, GU, H., SUENAGA, K. AND IIJIMA, S.	*Chem Phys. Lett*, 1997, **279**, 264–269
8.	SEN, R., GOVINDARAJ, A AND RAO, C. N. R.	*Chem Phys Lett.*, 1997, **267**, 276–280.
9.	SEN, R, GOVINDARAJ, A. AND RAO, C. N. R.	*Chem. Mater*, 1997, **9**, 2078–2083
10.	RAO, C. N R, SEN, R. SATISHKUMAR, B. C. AND GOVINDARAJ, A.	*Chem. Commun.*, 1998, 1525–1526
11.	SEN, R. *et al.*	*J. Mater. Chem.*, 1997, **7**, 2335–2337
12.	SEN, R. *et al.*	*Chem Phys. Lett.*, 1998, **287**, 671–676.
13.	SEN, R, SUMATHY, R. AND RAO, C. N. R.	*J. Mater. Res.*, 1996, **11**, 2961–2963.
14.	SEN, R., SUMATHY, R. AND RAO, C. N. R.	*J Mol. Struc. (Theochem)*, 1996, **361**, 211–216
15	SEN, R., SATISHKUMAR, B. C., RAINA, G. AND RAO, C N. R.	*Fullerene Sci. Technol*, 1997, **5**, 489–502.

Thesis Abstract (Ph.D.)

Some studies on interface states in GaAs MESFETs and HJFETs by V. R. Balakrishnan
Research supervisors: Profs Vikram Kumar and V. Venkataraman
Department: Physics

1. Introduction

The development of devices based on gallium arsenide, one of the most widely used III–V compound semiconductor materials, over the last 30 years, has attained maturity, due to ever-increasing applications in optoelectronics, microwave and millimeter wave devices and high-speed digital integrated circuits. GaAs metal semiconductor field effect transistors (MESFETs) and more recently, heterojunction field effect transistors (HJFETs) have been developed and used extensively for high-frequency low-noise applications. Rapid development, in device physics and the technology fronts, has resulted in an assortment of research studies dealing with the problems associated with non-ideal device performance. One of the intriguing problems which has not been fully explained deals with the semiconductor interfaces in GaAs devices. In the case of the MESFET, the semiconductor–insulator interface in the ungated regions (between source and gate and gate and drain) and the interface between the active layer and the semi-insulating substrate is known to cause several performance-related anomalies in the device behavior.[1] Similar problems have also been noticed in HJFETs where it has been attributed to the presence of bulk traps such as DX centres[2] and defect states existing at hetero-interfaces.[3] It is therefore highly probable that the ultimate success in removing such device-performance anomalies will eventually depend on the ability of the interface-processing technology, both the surface passivation in the case of MESFETs and optimum growth of dislocation-free hetero-interfaces in the case of HJFETs.

This work attempts to examine some of the device-performance anomalies which have persisted in GaAs MESFETs and HJFETs due to the role of deep-level defects present at the surface and interfaces of these devices. An effort is also made to understand the origin of these interface states and the physical mechanism by which they cause such device misbehavior. Finally, a model has also been developed to analyze the effect of interface states on the metal gate-interfacial layer-semiconductor Schottky contact in both of these devices.

2. Experimental

The devices used in this work are commercial ion-implanted microwave MESFET with 1 μm gate length and a pseudomorphic HJFET with gate dimensions 2.0×200 μm. The forward and reverse I–V characteristics at various temperatures were measured using the Keithley 428 current amplifier and a liquid nitrogen cryostat. Polaron 4600 DLTS system and 428 current amplifier were employed for the conductance deep-level transient spectroscopy (DLTS) measurements. The temperature dependence of transconductance at different frequencies was studied using the SRS 830 DSP dual-phase lock-in amplifier, the 428 current amplifier and the HP8116A function generator.

3. Results and discussion

The low-frequency transconductance dispersion in GaAs MESFETs as shown in Fig. 1 and the anomalous 'hole'-like peak observed in the conductance DLTS spectra in the case of MESFETs have been experimentally shown to be caused by a large concentration of surface states present in the ungated regions. The presence of an associated thin surface conduction channel in the ungated regions has also been experimentally verified by measuring the temperature-depend- ent reverse leakage current. The capture and emission processes from the surface states have been explained using a two-dimensional interface state band model.[4]

The effect of deep-level traps on some of the anomalous effects observed in a commercial pseudomorphic HJFET has also been experimentally examined. The reverse I–V characteris-

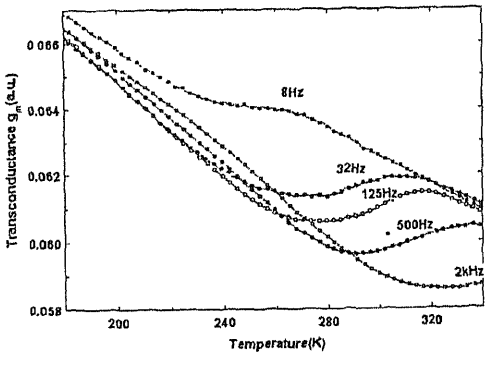

FIG. 1. Experimental transconductance vs temperature plots taken at different frequencies.

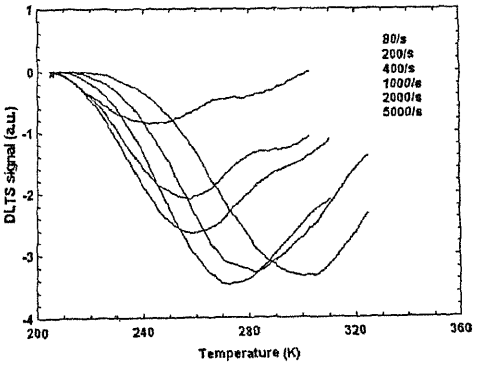

FIG. 2. Conductance DLTS spectra showing negative 'hole'-like peak at different emission rates. The voltages are *V(fill)* = +0.1 V, *V(reverse)* = –0.4 V. The fill pulse width is 10 ms.

tics and AC transconductance studies indicate the presence of interface electron traps near the hetero-interface. Conductance DLTS measurements performed both in the emission and capture modes have also demonstrated the presence of interface electron raps. Figure 2 shows the negative 'hole' like peak observed in the emission mode. This has been attributed to a strong change in the 2DEG mobility due to the change in the occupancy of interface electron traps.[5] The model developed to explain the conductance DLTS spectra was also effectively used to explain the low-frequency transconductance dispersion observed at different temperatures.

The departure from the ideal metal–semiconductor contact, in the case of a metal gate–interfacial layer–semiconductor Schottky contact in GaAs MESFETs and HJFETs has also been explained in this work. A non-equilibrium steady-state theory has been developed to determine the interface state density quantitatively between the metal and semiconductor from the experimental forward I–V characteristics carried out at different temperatures. The results show that a maximum value of ideality factor is observed in an intermediate value of applied forward bias.[6] This probably indicates the presence of a peak in the interface state density. It has also been found that the change in the barrier height due to trapping and detrapping of the carriers from the interface states between the metal and the semiconductor is independent of temperature above approximately 300 K.

References

1. BALAKRISHNAN, V. R , KUMAR, V. AND GHOSH, S — Experimental evidence of surface conduction contributing to transconductance dispersion in GaAs MESFET's, *IEEE Trans.* 1997, **ED-44**, 1060–1065
2. GHOSH, S. AND KUMAR, V. — Direct evidence for negative U nature of the DX centers in AlGaAs, *Phys. Rev. B*, 1992, **46**, 7533
3. HONG, W P., OH, J. E , BHATTACHARYA, P. K. AND TIWALD, T. E. — Interface states in modulation doped InAlAs/InGaAs heterostructures, *IEEE Trans.*, 1988, **ED-35**, 1585–1590.
4. HASAGAWA, H. AND SAWADA, T. — On the electrical properties of semiconductor interfaces in metal-insulator-semiconductor structures and possible origin of interface states, *Thin Solid Films*, 1983, **103**, 119–140
5. TAKIKAWA, M. — Electrical properties of interface traps in selectively doped AlGaAs/GaAs heterostructures, *Jap. J Appl Phys.*, 1987, **26**, 2026–2032.
6. MAEDA, K., IKOMA, H., SATO, K. AND ISHIDA, T. — Current–voltage characteristics and interface density of Schottky barrier, *Appl. Phys. Lett.*, 1993, **62**, 2560–2562.

Thesis Abstract (Ph.D.)

Near threshold fatigue crack growth and fracture toughness studies in zirconium, Zr-15%Ti and Zircaloy-2 by Azharul Haq

Research supervisors: Prof. E. S. Dwarakadasa and Dr. S. Banerjee (BARC)

Department: Metallurgy

1. Introduction

Fatigue crack growth (FCG) at low stress-intensity ranges is strongly influenced by microstructure because of the propensity of the cracks to grow along specific crystallographic planes resulting in crack tortuosity and crack closure, besides local crack tip retardation at boundaries.[1] Controlling microstructural unit (CMU) size, slip behaviour and presence of second-phase particles are believed to affect the near-threshold fatigue crack growth (NTFCG) behaviour by controlling the nature and extent of slip or twinning at the crack tip.[2] However, several aspects pertaining to the microstructure influence on NTFCG still remain unresolved.[3] These include identification of the real CMU among grain, packet or lath size; the role of P-free plate boundaries in deviating cracks; the existence of a bilinear FCG and the nature of relationship between the crack-tip plastic zone size (r_y) and the transition stress-intensity range (ΔKT);[4] effect of crystallographic orientation on fracture morphology specifically in anisotropic materials; the exact micromechanism of FCG particularly in acicular microstructures where the adjacent lamellae are either separated by small angle or twin-related boundaries, tile role of deformation twinning and intermetallic precipitates on crack path and FCG resistance, etc.[5] A systematic investigation was undertaken in order to gain an insight into these issues by examining in detail the effect of microstructure on NTFCG and fracture toughness (FT) in pure zirconium, Zr-15%Ti single-phase alloy and multicomponent multiphase commercial Zircalloy-2.

2. Experimental

The experimental steps were so designed as to develop various microstructures and involved tensile, FCG and fracture toughness testing in accordance with ASTM test procedures; detailed crack path and fractographic examination and transmission electron microscopy of deformation microstructure in zirconium. Aspects related to crack closure, which play a very signifi-

FIG 1 Photograph of a typical compact tension specimen.

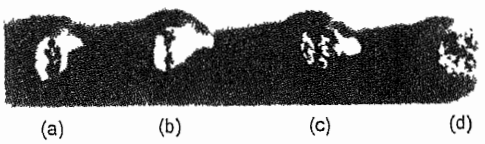

FIG 2 Fractured tensile specimens exhibiting variation in cross-sectional ellipticity with equiaxed α-zirconium grain size: (a) 8, (b) 15, (c) 35 and (d) 80 μm

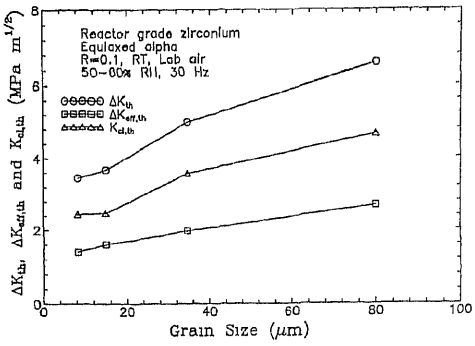

FIG. 3. Effect of equiaxed α-zirconium grain size on threshold fatigue properties.

cant role in NTFCG regime at low stress ratios, were given due consideration. All the compact tension specimens were metallographically prepared (Fig. 1).

3. Results

In the first phase, concerned with grain size effects, commercial purity zirconium was chosen to avoid the complicating factors generally present in multiphase commercial alloys. Different processing parameters employed in generating four different grain sizes resulted in significant variation in crystallographic texture. Although no texture analysis was carried out, the variation in texture with grain size was evident from the varying amounts of cross-sectional ellipticity exhibited by the fractured tensile specimen (Fig. 2). The influence of texture on static mechanical properties like tensile properties and ductile fracture toughness (J_{IC}) appeared to be dominating enough to mask the dependence of these properties on grain size. However, both applied and intrinsic threshold stress intensity ranges (ΔK_{th} and $\Delta K_{eff,\,th}$, respectively) as well as the closure stress intensity at threshold ($K_{cl,th}$) increased systematically with increasing grain size. ΔK_{th} increased more rapidly with grain size as compared to $\Delta K_{eff,\,th}$ (Fig. 3). The overriding effect of grain size on NTFCG behaviour could be understood in terms of crack deviation and deflection mechanisms leading to increased crack tortuosity with grain size. Fractography

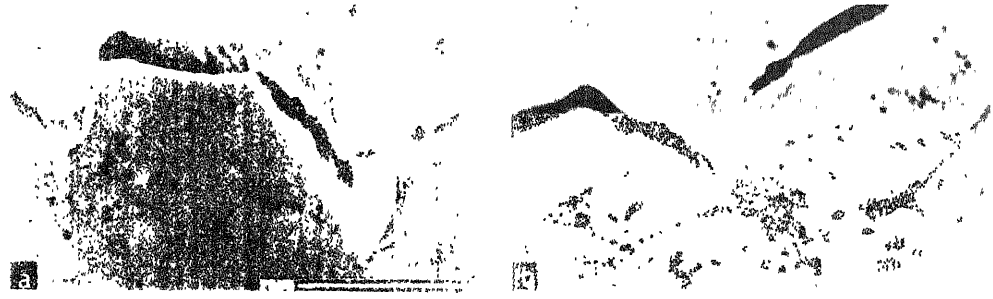

FIG. 4. Crack path profiles in equiaxed α-zirconium exhibiting (a) mode II and (b) mode III displacement in crack surfaces leading to the development of contact points.

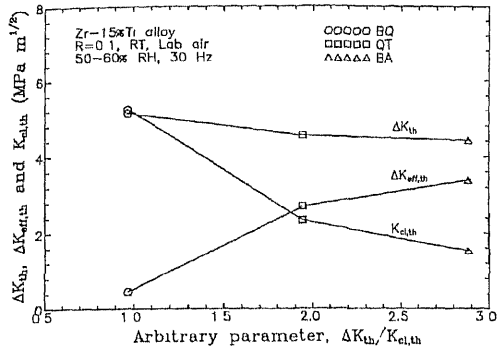

FIG 5 Effect of microstructure on threshold fatigue properties in Zr-15%Ti alloy

results suggest that slip reversibility also makes a significant contribution, which was further confirmed by the presence of planar dislocation arrays in the deformation microstructure. The crack closure mechanisms were identified as those induced mainly by roughness and to some extent by oxide. Contact points resulting from mode II (Fig. 4a) and mode III (Fig. 4b) displacement of crack surfaces were evident. Fracture surfaces corresponding to stages II and I were dominated by fissure striations and facets, respectively. However, no sharp transition was observed in either fracture surface morphology or FCG curve. Static fracture comprised dimpled rupture, the walls of large dimples exhibiting ripples.

The second phase of the programme consisted of investigating the influence of various β-transformed microstructures, viz., martensitic (BQ), tempered martensitic (QT), and Widmanstatten (BA) microstructures, on NTFCG behaviour in Zr-15%Ti alloy. This alloy is an ideal system for investigating microstructural influence not only because it allows various microstructures to be developed but also because the β to α transformation can be taken to completion thus avoiding possible complicating influence of retained β. The alloy picked up oxygen during hot forging operation, carried out under an apparently insufficient protective coating of borosilicate glass and hence exhibited high strength and low toughness. The hydrogen picked

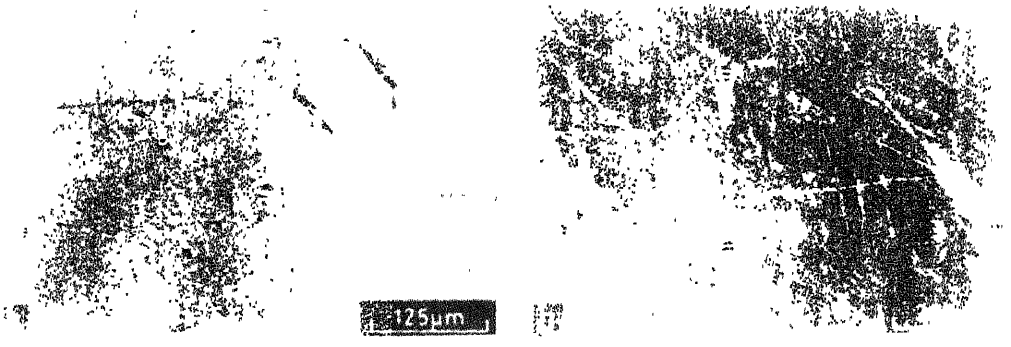

FIG. 6 Discontinuous crack growth in the BQ microstructure of Zr-15%Ti alloy at two different locations resulting in the development of intact ligaments (indicated by arrows in (b)).

FIG. 7. SEM fractographs of Zr-15%Ti alloy in the BQ condition exhibiting failure of intact ligaments by shear. A higher magnification view of the area, indicated by an arrow in (a), is shown in (b).

up possibly during quenching operation further impaired the toughness of the BQ and QT microstructures. All the three microstructures gave valid K_{IC} that decreased in the order of BA, QT and BQ microstructures. $\Delta K_{eff,\,th}$ and fracture surface roughness also decreased in the same order; however, $K_{eth,\,th}$ and ΔK_{th} surprinsingly increased (Fig. 5). A new crack closure mechanism involving crack bridging by shear ligaments (CBSL) was proposed and invoked to explain the above discrepancy. This inference was based on the observation of intact ligaments in the crack wake (Fig. 6) and shear ligaments (Fig. 7) on the fracture surface. CBSL was observed particularly in the BQ microstructure, which resulted in the K_{cl}/K_{max} value approaching unity. The β-free interplatelet boundaries were observed to effectively deviate cracks in contradiction with the widely accepted view of the presence of a thick and continuous β layer being considered essential for promoting crack deviation. Fractographic facets of dimensions, comparable to the plate size, further corroborated this observation. An interesting occurrence of ductile striations and faceted fracture in the alternate lamellae within a given colony suggested that the platelets were twin related. The fracture morphology appeared to be dependent mainly on the crystallographic orientation of a given CMU with respect to the stress axis.

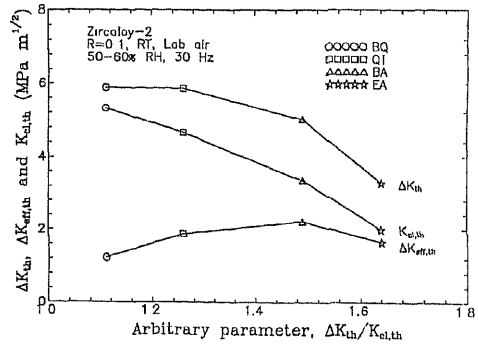

FIG. 8. Effect of microstructure on threshold fatigue properties in zircalloy-2.

Zircalloy-2 is widely used in the nuclear industry for various in-core reactor components. It is mainly a Zr–Sn alloy containing small amounts of Fe, Cr, Ni and oxygen. The microstructure comprises largely of a single phase containing intermetallic precipitates predominantly at various interfaces. The four microstructures developed in Zircalloy-2 as part of the third phase of the programme were BQ, QT, BA and equiaxed-α (EA). The microstructural influence on NTFCG behaviour followed the same trends as that observed in Zr-15Ti alloy, the EA microstructure exhibiting the lowest ΔK_{th} and $K_{cl,,th}$ (Fig. 8). The variation in $K_{cl,\,th}$ with microstructure could be attributed to RICC. J_{IC} decreased in the order of QT, EA, BQ and BA microstructure, which was largely rationalized in terms of the observed tensile toughness. The crystallographic cracks observed in the BQ microstructure were seen to cross even the prior β grain boundaries without any apparent deviation suggesting a possible orientation relationship between the prior β grains. Fatigue cracks exhibited propensity for growth along interfaces aided possibly by the presence of intermetallic precipitates.

4. Conclusion

Some of the observations common to the three materials investigated are as follows. At intermediate ΔK values (Paris regime) the fatigue crack path largely remained confined to mode I crack growth plane and exhibited duplex slip markings whereas at low ΔK values (near threshold regime) the crack path was predominantly crystallographic suggesting the operation of duplex slip mechanism. Accordingly, the fatigue fracture morphology largely comprised striations in the Paris regime and faceted (shear or cyclic cleavage) feature in tile near threshold regime. The contribution of crack closure exhibited an increasing trend with crack length. Twin and lath/plate boundaries offered greater resistance to crack growth through crack deflection mechanism when the crack had to cut across them. The acicular microstructures comprising large colonies and prior β grain sizes offered the best FCG resistance.

References

1. SURESH, S — *Fatigue of materials*, Cambridge Univ. Press, Cambridge, 1991.
2. RITCHIE, R. O. — *J. Engng Mater. Technol., Trans ASME*, 1977, **99**, 195.
3. RITCHIE, R. O. — *Int. Metall Rev*, 1979, **24**, 205
4. DICKSON, J I., BAILON, J. P. AND MASOUNAVE, J. — *Can. Metall. Q.*, 1981, **20**, 317.
5. PARRIS, P C. AND ERGODEN, F. — *J. Basic Engng*, 1963, **85**, 528.

Thesis Abstract (Ph.D.)

On development of intelligent tools for applications in energy control center by A. Narendranath Udupa
Research supervisors: Profs K. Parthasarathy and D. Thukaram
Department: Electrical Engineering

1. Introduction

Design, operation, and control of today's large-scale power systems with interconnections, sophisticated operational strategies, reliability and safety concerns, involve demanding computational issues. Computerized modern energy control centers (ECC), with varying degrees of sophistication, have evolved over the years to meet the above objectives. The growing complexity and sophistication of power systems have necessitated the development of intelligent tools such as artificial neural networks (ANNs), and fuzzy logic-based systems, to aid various planning and control operations in an ECC. This work addresses certain issues concerning application of ANN and fuzzy logic to aid the solution of various power system problems.

Many recent works have reported application of (i) ANNs for power systems such as load forecasting, power system stabilizer design, unit commitment, security assessment, economic load dispatch, and fault analysis, and (ii) fuzzy logic for power systems such as system control, optimization, diagnosis, information processing, decision support, system analysis and planning. Five areas in power system analysis have been selected in this work for application of intelligent tools. ANNs have been applied for two of them and fuzzy logic for the remaining. The algorithms developed for the selected key problems are:

1. ANN-based methods for the determination of fault location and fault resistance on transmission lines.
2. ANN for static voltage stability assessment.
3. Fuzzy logic-based controller for voltage stability enhancement.
4. Fuzzy logic-based controller for voltage profile improvement.
5. Fuzzy logic-based controller for network overload alleviation.

The algorithms developed are tested on various sizes of power networks including practical systems of 24 bus, 82 bus, and 24-node EHV, and a modified IEEE 30 bus sample system.

2. ANN-based algorithms

2.1. *Fault location algorithm*

The fault location algorithm[1] is a key element in the digital relay for transmission line protection. The potential applicability of ANN techniques for determination of fault location and fault resistance on EHV transmission lines with remote end in-feed is analyzed. A methodology where ANN is sought to achieve the desired accurate solution with easily measurable quantities being presented as inputs is proposed in this work. Most of the ANN applications make use of the conventional multilayer perceptron (MLP) model based on back propagation algorithm. But this model suffers from the problem of slow learning rate. A modified cascade correlation-based ANN learning technique[2] has been proposed for the determination of fault location and fault resistance. A reasonably small NN is built automatically without guessing the size, depth and connectivity pattern of the NN in advance. An analytical method considering all the parallel power corridors across the line under consideration is also developed. Results of simulation studies performed on a 400 kV transmission line are presented for illustration purposes. The performance of the proposed ANN is compared with the analytical algorithms and conventional MLP algorithm for different combinations of pre-fault loading condition, fault resistance and fault location. Figure 1 shows the fault location and fault resistance

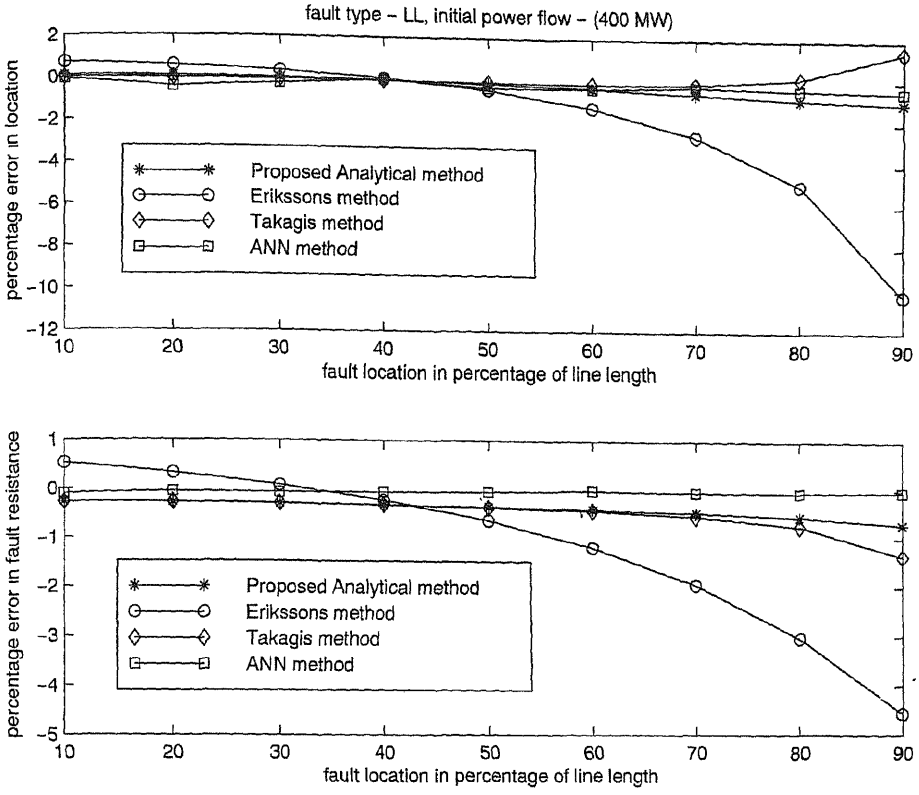

FIG. 1. LL fault

determined by various analytical techniques such as Ericksson method,[1] Takagis method and the developed ANN method. The results of the proposed analytical method and ANN methods are found to be more accurate.

2.2. *Voltage stability assessment algorithm*

Voltage collapse phenomena has been observed in many countries and has been analyzed extensively in recent years. Unavailability of sufficient reactive power sources to maintain normal voltage profiles at heavily loaded buses are the prime reasons for the voltage collapse.[2] In some cases, because of load variations, voltage profiles may not show abnormality prior to voltage collapse and the system operators may not get warning of it. It is important to assess system voltage stability and use the reactive power sources judiciously to improve the voltage stability of the system. Methods which can provide fast assessment of system voltage stability and optimum controls of reactive power sources would be useful for online application. An application of ANN where the aim is to achieve fast voltage stability margin assessment of power network in ECC, with reduced number of appropriate inputs to ANN, is presented in this work. L-index, which is a function of network elements, and generator/load bus voltages,

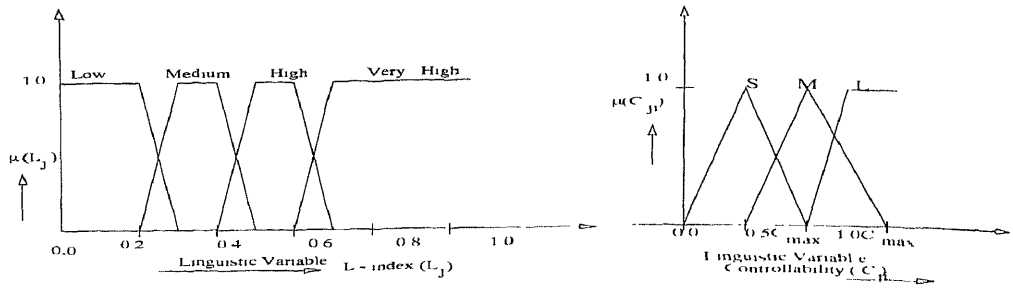

FIG 2 Membership functions of linguistic terms

has been used for assessing the voltage stability margin. An improved algorithm for static voltage collapse proximity indicator (L-index), based on operational load flow (OLF), is developed for the purpose of generating input–output pairs. Inputs play an important role in ANN learning and hence meaningful training patterns in the normal operating range of the system are to be generated. For this purpose, an algorithm for reactive power optimization using LP technique to enhance voltage stability is proposed. The objective is to maximize the voltage stability or minimize the sum of L-indices (min $\sum_j L_j$). Training patterns for the ANN are generated by running the developed LP technique for reactive power optimization. Input parameters are chosen from the available measurements based on:

- statistical correlation process,
- sensitivity matrix approach,
- contingency ranking approach, and
- concentric relaxation method.

Investigations are carried out on the influence of information encompassed in the input vector and the target output vector, on the learning time and test performance of MLP-based ANN model.

3. Fuzzy logic-based algorithms

3.1. *Fuzzy logic-based voltage control*

Two areas are selected for fuzzy logic-based voltage control.

- voltage stability enhancement,
- voltage profile improvement.

The voltage stability enhancement approach translates voltage stability index and controlling variables into fuzzy set notations in order to formulate the relation between voltage stability level and controlling ability of controlling devices.[3] The control variables considered are:

- switchable VAR compensators,
- OLTC transformers, and
- generator excitations.

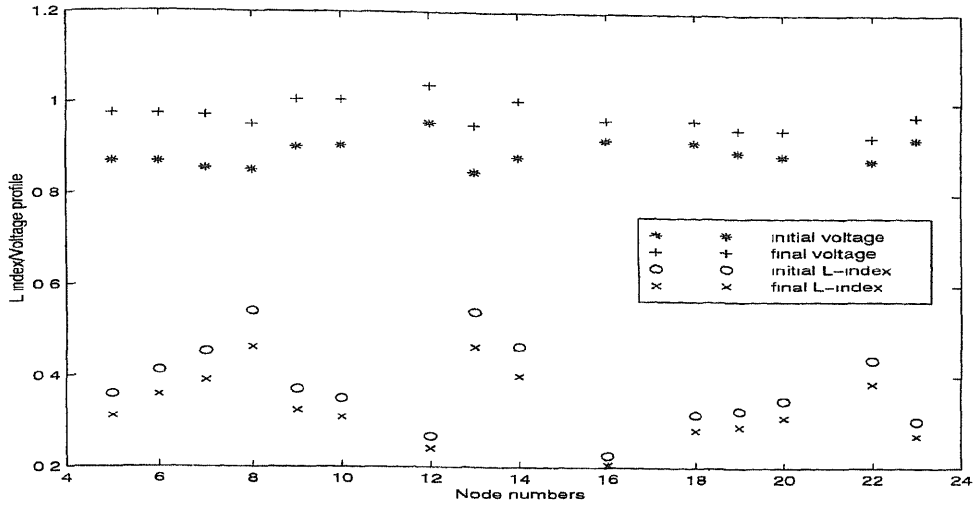

FIG. 3. Base case: 24-bus system, values of L index and voltage profile at critical nodes

This approach consists of 12 rules defining 4 terms (very high, high, medium, and low) for voltage stability index and 3 terms (large, medium, small) for controllability of controllers. Controllability indicates the sensitivity of the voltage stability index to controllers and the controller margin available. The sensitivity of the voltage stability index to controllers is determined in two steps. In the first step, the sensitivity of the voltage stability index to voltage deviation is obtained and in the second the sensitivity of the voltage deviation to controllers is

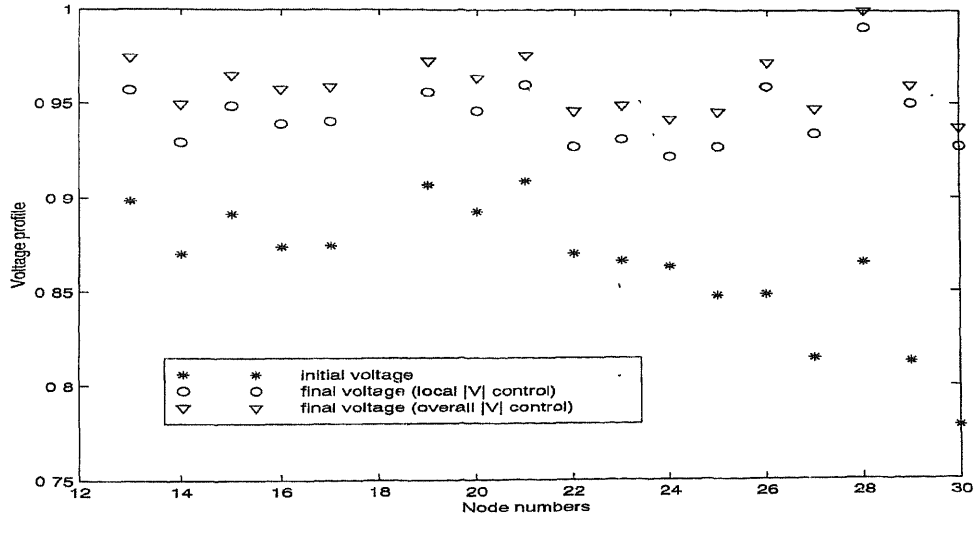

FIG. 4. Base case: 30-bus system, voltage profile at critical nodes.

determined. Figure 2 shows the membership functions of linguistic terms used and Fig. 3 the L indices and voltage profile at critical nodes of a 24-bus system before and after the implementation of the proposed fuzzy control.

The voltage profile improvement[4] approach translates voltage deviation and controlling variables into fuzzy set notations. Controllers are selected based on two criteria. One, local controllability of a controller towards a bus having poor voltage profile and two, overall controllability of a controller towards all the buses having unacceptable voltage profile. Controllability of a controller in reducing the voltage deviation at only the given bus is considered in local controllability approach. In this approach, the effect of the selected controller on the remaining buses is ignored. Also, a new concept of overall improvement in voltage profile is proposed which not only considers the reduction of voltage deviation at a given bus but also takes into consideration the effects of suggested controller action on the remaining buses, based on sensitivity. Voltage deviation and controlling variables are translated into fuzzy set notations to formulate the relation between voltage deviation and controlling ability of controlling devices. Six terms (negative large, negative medium, negative small, positive small, positive medium, and positive large) each for defining the voltage deviation and controllability of controller are specified. Controllability indicates both the sensitivity of the voltage deviation to controllers and controller margin available. Sensitivities of the voltage deviations are determined in only one step, thus reducing the computational burden. A fuzzy rule-based system consisting of 36 rules is formed to select the controllers, their movement direction and step size. The performance of the two fuzzy logic-based systems, one for voltage stability enhancement and the other for voltage profile improvement, is compared. It is seen that methodologies based on overall voltage profile enhancement and voltage stability improvement give nearly equal performances, better than that of local controllability approach for voltage profile enhancement. It is also observed that the computational burden is less in the methodology based on overall voltage profile enhancement while forming the sensitivity matrix. Hence this

Table I
82-Bus system; base case

Pre-rescheduled line flows (MVA)	
Overloaded lines	$L_{25} = 120\ 5$, $L_{26} = 506.3$, $L_{36} = 603.2$
Fully loaded lines	$L_2 = 96\ 13$, $L_3 = 272.8$, $L_{45} = 272.1$, $L_{46} = 292.4$, $L_{47} = 293.0$, $L_{50} = 139.2$
V_{min}	$V_{81} = 0.900$
P_{loss}(MW)	179.1
Post-rescheduled line flows (MVA)	
Overloaded lines	$L_{26} = 481\ 7$
Fully loaded lines	$L_3 = 284.4$, $L_{25} = 114.7$, $L_{36} = 586.2$, $L_{45} = 271.3$, $L_{46} = 282.2$, $L_{47} = 293.0$, $L_{50} = 138.6$
V_{min}	$V_{81} = 0.902$
P_{loss}(MW)	174.4
Generation rescheduling in MW	$P_4^+ = 40.0, P_9^+ = 44.0, P_{10}^+ = 40.0$
	$P_2^- = 44.6, P_3^- = 33.8, P_5^- = 27.0$
	$P_6^- = 10.4, P_7^- = 8.3$

approach is preferable for online application in ECC. Figure 4 indicates the voltage profile obtained at critical nodes of IEEE 30 bus system because of the proposed fuzzy control.

3.2. *Network overload alleviation algorithm*

The fuzzy control approach is extended for network overload alleviation in power networks by active power generation rescheduling.[5] Accurate generation shift sensitivity factors (GSSF) are computed using more realistic OLF model. Overloading of lines and sensitivity of controlling variables are translated into fuzzy set notations to formulate the relation between overloading of line and controlling ability of generation scheduling. A fuzzy rule-based system is formed to select the generator, and the amount of generation rescheduling at the selected generator. Also, overall sensitivity of line loading to each of the generation is considered for the rescheduling process. Table I shows the results obtained on an 82-bus Indian power network.

4. Conclusions

Studies were carried out for various practical Indian power networks under simulated conditions. Fault-locating algorithm was tested on a 400 kV, 300 km, double circuit line under simulated conditions. Results of the other developed algorithms obtained for a modified IEEE 30-bus system and three Indian power networks of 24-bus, 82-bus, and 24-node EHV are presented for illustration purposes. Comparing the results obtained by conventional techniques with those of proposed algorithms, it can be concluded that the proposed algorithms, ANNs and fuzzy systems, give faster and acceptable solutions. The intelligent tools and analytical techniques are thus found to be suitable for implementation in ECC as a decision aid to the operator and for online control of power systems.

References

1. ERIKSSON, L., SAHA, M. M AND ROCKEFELLER, G. D — An accurate fault locator with compensation for apparent reactance resulting from remote-end feed, *IEEE Trans.*, 1985, **PAS-104**, 424–436.
2. EL-KEIB, A A. AND MA, X. — Application of artificial neural networks in voltage stability assessment, *IEEE Trans* , 1995, **PS-10**, 1890–1896.
3. SU, C-T AND LIN, C-T — A new fuzzy control approach to voltage profile enhancement for power systems, *IEEE Trans.*, 1996, **PS-11**, 1654–1659.
4. KIRSCHEN, D. S. AND VAN MEETEREN, H P. — Mw/voltage control in a linear programming based optimal power flow, *IEEE Trans.*, 1988, **PS-3**, 481–489.
5. BANSILAL, THUKARAM, D. AND PARTHASARATHY, K. — An expert system for alleviatioin of network overloads, *Electric Power System Res.*, 1997, **40**, 143–153.

Thesis Abstract (Ph.D.)

Soundness and completeness results in partially interpreted logics by K. Suman Roy
Research supervisor: Prof. V. Chandru
Department: Computer Science and Automation

1. Introduction

Most applications of logic in computer science call for a priori interpretation of some function and predicate symbols over predefined computational domains. This is called a *pre-interpretation* or a *partial interpretation* of logic. For a given formula, we can check the validity of the formula assuming the pre-interpretation. Many practical problems in different areas of computer science have been successfully modeled using partially-interpreted logics.

The *Constraint Logic Programming (CLP) scheme* was introduced by Jaffar and Lassez.[1] CLP began as a natural merger of two declarative paradigms: constraint solving and logic programming. This combination helps make CLP both expressive and flexible, and in some cases, more efficient than other kinds of programs. Although CLP is a relatively new field, its applications have been successfully developed in several directions ranging from databases, options trading, DNA sequencing to job-shop scheduling. The scheme of CLP gives a formal framework, based on constraints, for the basic operational, logical and fixpoint semantics of an extended class of logic programs. This framework extends traditional logic programming in a natural way by treating unification in logic programming as constraint solving. Further, the CLP scheme integrates constraint solvers in any predefined computational domain into the SLD resolution method.

In many real-world applications it is needed to write logic programs which require the expression of negation or intention. To this end, the designers of Prolog added many nonlogical extensions to the Horn-clause kernel, notable among them is the *Negation as (Finite) Failure* rule which facilitates the derivation of negative facts. In the framework of CLP we give a resolution proof of the soundness and completeness of the Negation as failure rule.

Embedded computer systems are widely used nowadays, examples of which include sensors, actuators, control circuits, etc. In these applications a computer has to interact with these devices. In the process, there are stringent real-time constraints on the computer response. Several real-time logics have been proposed to reason about these systems. In these logics, properties of time-dependent behaviour are expressed as formulas and the logical consequences of any given specification are derivable. In real-time logics people consider the natural numbers $\mathbb{N}$ or real numbers $\mathbb{R}$ as the underlying time domain. Both $\mathbb{N}$ and $\mathbb{R}$ have binary operators + (addition) and * (multiplication) defined on them. In these logical languages those symbols are assumed to be preinterpreted. Since $\mathbb{N}$ and $\mathbb{R}$ are well-ordered time frames they can be ordered by the pre-interpreted predicate symbol $\leq$. In part two of the work we investigate a real-time logic called *Neighbourhood Logic* (NL). We then extend NL to another real-time logic called *Duration Calculus* with reals as the underlying time domain and design a proof system for it. This proof system is shown to be sound and relatively complete with respect to Neighbourhood Logic.

2. Negation as Failure in logic programming

Logic programming evolved in the early 1970s as a direct outgrowth of earlier work in automated theorem proving and artificial intelligence. The major breakthrough in automated deduction was provided by the landmark paper of Robinson[2], in which he introduced *resolution* as an inference rule. Subsequently, people came up with many refutation procedures which were refinements of the original resolution rule. The most notable refutation procedure

were refinements of the original resolution rule. The most notable refutation procedure adopted in logic programming is due to Kowalski, called *SLD resolution*[3, 4] and is applied on Horn clauses.

SLD resolution[3, 4] derives only positive consequences (namely, conjunctions of atoms) of Definite logic programs. However, in many circumstances, it is also useful to derive negative consequences. A classic example of usefulness of negative consequences is that of a timetable problem which connects explicitly, but the absence of connections implicitly.

The derivation of negative facts from Definite logic programs was facilitated by Clark[5], who introduced Negation as (Finite) Failure rule, which states that $\neg A$ is a consequence of a program P if a finitely failed SLD-tree of $P \cup \{\neg A\}$ exists (in short, if A fails), where A is a *ground atom*. Clark also introduced the notion of complete logic programs[5] which strengthen the program by interpreting implications as equivalences. That is, a Definite program is completed by transforming the 'if–then' statement into an 'if and only if' statement. He proved the soundness of this rule, that is, if the completion of a Definite program entails $\neg A$, then the program (finitely) fails to SLD-prove A, where A is an atom (without any variable). Jaffar and Lassez[1] proved the completeness of the negation as failure rule with respect to Clark's semantics. They showed that given a definite clause program if a negative goal is entailed by the completion of the program then the program finitely fails to prove the goal. This proof was later simplified by Wolfram *et al.*[6]

None of the above-mentioned completeness proofs are procedural. Jaffar *et al.*[7] proposed the following approach towards a resolution proof. They pointed out that the completion of a logic program P[3,4] consists of two distinct pieces:

- the If part which contains the clauses in P; this can be obtained by replacing the biconditionals in the completed program by ifs.
- the OnlyIf part which is obtained by replacing the biconditionals in the completed programs by only ifs.

Further, they argued that for propositional programs the completion entails a negated atom if and only if the OnlyIf part entails it. Our goal is to develop this idea in full, for the predicate and constraint logic programming cases, and thus complete the program sketched by Jaffar *et al.*[7] We will work with a variant of OnlyIf(P) which will be called FI(P). Our proof will be constructive in the following sense: it will indicate how a finite failure SLD-tree for a goal G and program P can be converted into a (resolution) proof of $\neg G$ from FI(P) and conversely. We also believe that our approach has finally led to a simple and comprehensible proof of the negation as failure rule.

The soundness and completeness theorems in the paper are proved via the following steps*:

1. We prove that a negative atom is entailed by the completion of a program P if and only if it is entailed by the FI(P) part.
2. We define a rule—$\forall$-SLD resolution—for proving negated goals from the FI(P). We prove it to be sound and complete for the usual first-order semantics.

3. We show that the following are equivalent: given a definite goal G, a definite program P and a satisfaction complete constraint theory τ

   - there is a $\forall$-SLD derivation of $\neg G$ from the FI(P) of the program completion.
   - P finitely fails to prove G.

3. Neighbourhood logic and duration calculus

Embedded systems are increasingly used in applications where they interact with physical processes. In these applications, they have to react to events within a prescribed time interval to produce an output before a certain delay has elapsed. In order to reason about them, numerical as well as symbolic time events have to be considered. *Thermostat* is an example of such system. The temperature of a room is controlled through a thermostat, which continuously senses the temperature and turns a heater *on* and *off*. The temperature is governed by differential equations. When the heater is off the temperature decreases according to some exponential function. When the heater is on the temperature rises, again according to some exponential function. We wish to keep the temperature between say, m and M degrees. The computer will be able to detect that and send instructions to turn the heater on and off.

In order to analyze such systems various real-time logics have been proposed. Some of them are temporal in nature where formulas are interpreted as instantaneous situations over discrete time points.[8] Other formalisms adopt different semantics and interpret formulas over intervals of time.[9–11] Among such interval logics, Interval Temporal Logic (ITL)[10] and more specifically, the Duration Calculus (DC)[11] are quite popular. DC is an extension of ITL in the sense that the temporal variables are written in the form of integrals of state variables.

ITL is a first-order modal logic which uses a binary modal operator "$^\frown$" (chop) which is interpreted as the operation of 'chopping' an interval into two parts. A modality is called *contracting*, if we can access only subintervals within an arbitrary given interval with this modality. The modality $^\frown$ is an example of a contracting modality. A shortcoming of ITL is that it is not expressive enough to be able to capture statements regarding the 'safety' properties of embedded systems since these properties depend on the behaviour on intervals outside the observed interval of time.

Another limitation of these logics is that when they are used in the specification of hybrid systems, the concepts of calculus such as limit, continuity and differentiability cannot be suitably formalized in them. These notions are neighbourhood properties of an interval which cannot be defined in these logics.

Chaochen and Hansen[12] have introduced a first-order logic for intervals called NL which has constructs for formalization of safety properties[13] as well as some important concepts in calculus. This logic has two expanding modalities, $\Diamond_l$ and $\Diamond_r$, called the *left* and *right neighbourhood modalities*, respectively. They also proved the adequacy of the neighbourhood modalities by deriving other unary and binary modalities from them.[12] Further, they extend NL

*A preliminary version of this work appeared in Chandru V., Roy, S. and Ramesh, S., Constructive Negation in Definite Logic Programs, *Proc. 2nd Asian Computing Science Conf.*, ASIAN'96, pp. 335–336, LNCS 1179, Springer, 1996.

to another real-time logic called Duration Calculus* by expressing a temporal variable as an integral of a state variable. In DC/NL they specify the *safety* properties of computing systems. Thus DC/NL provides an environment in which we can specify and reason about embedded systems.

In this work, we present a thorough investigation of NL and DC/NL. Extending the proof system of NL, we introduce a proof system for DC/NL including two induction rules. We prove the soundness and relative completeness of this proof system for DC/NL.**

References

1. JAFFAR, J. AND LASSEZ, J-L. — *Constraint logic programming*, Technical Report 86/73, Department of Computer Science, Monash University, 1986.
2. ROBINSON, J. A. — A machine-oriented logic based on the resolution principle, *J. ACM*, 1965, **12**, 23–41.
3 APT, K. R — *Logic programming*, *Handbook of theoretical computer science*, (J. van Leeuwen, ed.), 1990, Elsevier, Vol. B, pp. 493–574.
4. LLOYD, J. W. — *Foundations of logic programming*, 2nd edition, Springer Verlag, 1987.
5. CLARK, K. L. — *Negation as Failure*, in *Logic and databases* (H. Gallaire and J. Minker, eds), Plenum Press, 1978, pp. 293–322.
6 WOLFRAM, D , MAHER, M. AND LASSEZ, J.-L. — *A unified treatment of resolution strategies for logic programs*, *Proc. 2nd Int. Conf. on Logic Programming*, 1984, pp. 263–276.
7 JAFFAR, J , LASSEZ, J-L. AND MAHER, M. J — *Some issues and trends in the semantics of logic programming*, *Proc. 3rd Int. Conf. on Logic Programming*, LNCS 225, Springer-Verlag, 1986, pp. 223–241.
8. ALUR, R. AND HENZINGER, T. A. — Real-time logics: Complexity and expressiveness, *Inf. Computation*, 1993, **104**, 35–77.
9 HALPERN, J. AND SHOHAM, Y — A propositional modal logic of time intervals, *Proc. First IEEE Symp on Logic in Computer Science*, Computer Society Press, 1986, pp. 279–292.
10. MOSZKOWSKI, B. — A temporal logic for multilevel reasoning about hardware, *IEEE Computer*, 1985, **18**(2), 10–19.
11. CHAOCHEN, Z., HOARE, C. A. R. AND RAVN, A. P. — A calculus of durations, *Inf. Processing Lett.*, 1991, **40**, 269–276.
12. CHAOCHEN, Z. AND HANSEN, M H. — *An adequate first order interval logic*, UNU/IIST Report No. 91, Revised report, December 1996.
13 SKAKKEBÆK, J. U. — Liveness and fairness in duration calculus, *CONCUR'94: Concurrency theory*, LNCS 836 (B. Jonsson and J. Parrow, eds), pp. 283–298, Springer-Verlag.

*We call it DC/NL to distinguish from the original Duration Calculus which was based on ITL which we call DC/IL

**These results have appeared in Suman Roy and Zhou Chaochen: *Notes in neighbourhood logic*, UNU/IIST Report, No. 97, 1997.

Thesis Abstract (M.Sc. (Engng))

Biobeneficiation of bauxite using *Bacillus polymyxa:* Investigations on calcium removal by S. S. Vasan
Research supervisors: Profs Jayant M. Modak and K. A. Natarajan
Department: Chemical Engineering

1. Introduction

India has abundant reserves of bauxite ($Al_2O_3.xH_2O$)[1], most of which are present as lean-grade (<50% Al) ore with impurities like calcium ($CaCO_3$) and iron (Fe_2O_3). In order to use this lean-grade bauxite for abrasive and refractory applications, it is required to remove the impurities to about 0.5% calcium (expressed as %CaO) and 1% iron (expressed as $\%Fe_2O_3$). Conventional pretreatment strategies have serious disadvantages and are not suitable for beneficiation of Indian bauxite, so India is currently importing alumina-based refractories from China in spite of having abundant resources. A biotechnological solution to this problem promises to be efficient, environment-friendly and economically viable. This work is part of an ongoing project to understand biobeneficiation of bauxite and develop suitable technology for the sponsors M/s Orient Abrasives Limited, New Delhi. It deals mainly with the investigations on calcium removal using a bacterium called (*Paeni*)*Bacillus polymyxa*.

2. Experimental and discussion

The Bromfield medium and its recent modifications[2] are not satisfactory for growing *Bacillus polymyxa* in an industrial scale. In this work, a cheaper, cane sugar-based formulation called 'ISF-2' is developed, and the technique of *overpopulation* is proposed to obviate the need for stringent aseptic culture techniques. In order to combat occasional contamination by fungi and yeast, a selective formulation based on azole fungicides is also tested. The growth of *Bacillus polymyxa* is characterized by an increase in cell number, consumption of sucrose, production of organic acid metabolites (and exopolysaccharides), and an increase in the acidity of the medium. A 3-day-old culture has cells in the order of 10^9, a pH of 3 or less due to production of organic acid metabolites, and can typically solubilise around 300 ppm as Ca. Preliminary experiments give evidence that *Bacillus polymyxa* is able to selectively remove calcium, iron and silica from bauxite, which makes it an ideal choice for biobeneficiation.

A column bioreactor is designed to carry out biobeneficiation in a fluidized bed as well as slurry configuration. With this reactor it is possible to achieve uniform mixing without agitation, and also aeration without air compression.[3] It can be operated in periodic cycles of fluidization mode (referred to as *flood* cycle) and stationary submerged mode (referred to as *drain* cycle) using a timer. Fluidization carried out intermittently (as *flood/drain* cycles) on particles of size –4/+5 mesh gives evidence to confirm our findings from preliminary experiments. Several operating considerations in the industry favour particle size in the range of –200/+300 mesh, so further experiments are carried out with particles of this size. As fine-sized particles are best processed as slurry, all our demonstration-scale experiments are carried out in a total recycle plug flow configuration that approximates a slurry reactor. A 0.25-HP centripetal pump is used to pump the culture solution from a storage tank. The leaching efficiency of calcium is

very good when we cascade using a 36-h culture, with a contact time of 24 h per cascade. The number of cascades required for bringing down calcium (from 2.8% to 0.5% CaO) in 5% pulp density of bauxite ore is three or four, depending on the nature of the culture. There is only around 20% removal of iron, and the efficiency of iron reduction is enhanced under anaerobic conditions. On the other hand, calcium removal is not affected much under anaerobic conditions, albeit aeration of the reactor helps. The temperature rise in the reactor is less than 10°C, so no cooling systems are required.

The mechanism and kinetics of calcium removal are investigated in detail. The studies show that the dissolution of calcium from the ore in a 3-day old culture (pH 2–3) is kinetically very fast, and there is almost an instantaneous neutralization of the culture to pH 5–6. This initial phase takes less than 1 h and corresponds to a sudden removal of calcium to saturation solubility levels. The bacterium also attaches itself to the ore within 20 min and then brings about weathering and dissolution of more calcium because of further metabolic activity. However, this second phase is slower and takes nearly 24 h, after which there is a slow decrease in the calcium solubility of the metabolite leading to some amount of re-precipitation of leached calcium. The saturation solubility of calcium is found to be a function of the extent of cell growth and the increase in saturation solubility correlates well with the growth curve of the organism. As it is found that the initial phase (I) of biobeneficiation corresponds to the action of metabolites in solubilising considerable amount of calcium, we can consider the indirect leaching (through metabolites) to be the most important mechanism in our case. The indirect leaching mechanism comprises the action of H^+ ions in releasing Ca^{++} from the ore matrix (acidolysis), and the nature of the organic acid to form water-soluble Ca complexes with its anion (complexation).

As the dissolution of accessible calcium is kinetically fast, it is possible to speed up the leaching operation by cascading the required number of times for a brief period of time (say 10 min) called *pulses*. The *expected number* of cascades that will leach 20% pulp density of the ore to the target (0.5% CaO) are 14 and 19 for metabolites with and without cells, respectively. The background leaching is compared by leaching with HCl and H_2SO_4 at the same pH, and it is seen that these require 25 and 45 cascades, respectively. This is because mineral acids are stronger and hence present in fewer molar levels compared to organic acids at the same pH.

Fundamental studies on calcium solubility are carried out to understand the science behind biobeneficiation, and there is a good agreement with the experimental observations. Biobeneficiation is a complex process comprising many sub-processes occurring simultaneously. A model is proposed describing these sub-processes using Monod's growth kinetics, Langmuir-type attachment of cells to the ore, growth-associated sucrose consumption and acid production, and calcium removal driven by the concentration gradient in the system. Kinetic parameters describing these sub-processes are independently estimated. The final model integrates the simultaneous effects of all these sub-processes and its predictions of the experimental trends are reasonably good for a first generation model.

References

1. Suss, A. G., Kovalenko, P. and Nandi, A. K. Some geological, mineralogical and technological features of Gujarat Bauxites, India, 1998 TMS Annual Meeting Technical Program Insert, November 1997, **JOM-47**, No. 11, 38.

2 ANAND, P., MODAK, J. M. AND NATARAJAN, K. A. Biobeneficiation of bauxite using Bacillus polymyxa: Calcium and Iron removal, *Int. J. Mineral Processing*, 1996, **48**, 51–60.

3. ANDREWS, G. F., NOAH, K. S, GLENN, A. W. AND STEVENS, C. J Combined physical/microbial beneficiation of coal using the flood/drain bioreactor, *Fuel Processing Technol.*, 1994, **40**, 283–296.

Thesis Abstract (M.Sc.(Engng))

Adaptive hierarchical RAID by Nitin Muppalaneni
Research supervisor: Prof. K. Gopinath
Department: Computer Science and Automation

1. Data storage becoming critical

With increasing amount of data being managed by the use of computers, the speed and reliability of data storage have become critical. The following are some of the trends that are driving the design of storage systems to higher level of availability and performance.

Ever widening speed gap between CPU and disks: Disk speeds have been increasing very slowly as compared to the CPU speeds. This is mainly due to the mechanical components in disks.

Increasing disk capacities: Advances in magnetic storage have led to the capacities of disks grow impressively. Commodity disks of 9 GB capacity are now available and may become common by 1998.

This makes reliability a major concern. The amount of time required to restore the data from backup (assuming that they exist) in the event of a disk crash will only increase and also restoring from backup involves human intervention and subsequent possibility of human error.

Advent of I/O intensive applications: With more and more computers being employed to access, manage and visualize data (or information), the applications employed will require very high data bandwidths. The data set sizes are also increasing. For these applications data storage performance becomes a bottle neck.

Lost data is lost business: As computers are increasingly employed in business, unavailability of data leads to lost business. This reason alone can justify extra cost in making the storage highly available.

2. RAID (Redundant arrays of inexpensive disks)

Redundant arrays of inexpensive disks or RAID is a popular method of improving the reliability and performance of disk storage by introducing redundancy into the storage system.

Of various levels of RAID, *mirrored* or RAID1 and *rotating parity* or RAID5 configurations have become most popular.

Mirroring or RAID1: In this scheme, disks are grouped into mirrored pairs. Two copies of the same data are maintained on each of the disks in a mirrored pair. A write is applied to

both the disks, whereas a read can be satisfied by either of the disks. This scheme provides the best performance and reliability, but incurs 100 per cent storage overhead. In degraded mode (when one of the disks in a mirrored pair fails), writes and reads are both serviced by the remaining disk.

Block interleaved parity or RAID5: The data is striped across the disks with each stripe of *N* stripe units containing *(N-1)* data stripe units and one parity stripe unit containing the parity of these data stripe units. The parity itself is distributed over disks, so that the parity updates do not become a bottleneck.

The given I/O request is first divided into sub-requests at the granularity of stripes. Each of these sub-requests is then handled independently. When all these sub-requests complete, the I/O is signaled as completed. Now let's look at how these stripe I/Os (or substripe I/Os) are handled. In *normal* mode (when all disks are functioning), a read is served in a way similar to RAID0. For a write that spans the entire stripe (called *full stripe write*), the parity is computed from the data and both the data and the parity are written. A partial write, on the other hand, requires reading the old parity/data, recomputing the parity and writing the data and parity; this procedure is referred to as *read-modify-write* cycle. This makes partial RAID5 writes doubly slower than reads or full-stripe writes. The performance of reads drops drastically when one of the disks fails and the storage enters *degraded mode*, as all reads to data from the failed disk need to be reconstructed from the data on other disks and the parity.

Also configuring RAID5 is an involved job. If the stripe size selected is too big, then most of the accesses become partial-stripe accesses. In addition to this, updates to the same RAID5 stripe are serialized as all of them need to update a common parity area, resulting in lower I/O parallelism. On the other hand, if the stripe is made too small the number of I/Os required to service a request increase resulting in longer queuing delays leading to poorer performance.

Comparing RAID1 and RAID5: Table I compares the *normal* mode performance and capacity overhead of RAID0, RAID1 and RAID5. The values are first-order approximations for the configuration with each disk connected to the host through a separate HBA. The actual values depend on seek distance, head synchronization, access patterns, etc. Also, this kind of configurations are rare to find; typical configurations contain multiple disks connected to one HBA. In such configurations, delays at the HBA and SCSI bus also need to be considered. It is clear that RAID1 offers better performance than RAID5, but with a 100% storage overhead.

Table I
Performance of RAID levels relative to that of RAID0

RAID level	Large accesses		Small accesses		Capacity overhead(%)	Max concurrency
	Read	Write	Read	Write		
0	100	100	100	100	0	N
1	100+	100–	100+	100–	100	N
5	100	100	100	50–	100/N	N

3. The problem and a solution

In the previous section we have compared RAID1 and RAID5. In this section, we argue for a better price-performance RAID. Then a case is made for hierarchical RAID.

3.1. *Problem: Need for better price-performance RAID*

The cost of a storage system mainly contains two components:

1. System cost: This is the cost of the system including disks, upgrades, etc.
2. Management cost: This is the cost of managing the system. Administrator's salaries, etc. come under this.

As the storage becomes increasingly complex, the cost of managing the system dominates the total cost. So a better price-performance RAID storage should not only require less initial investment (unlike RAID1) but also less managing.

3.2. *A solution: Adaptive hierarchical RAID*

A solution to the above problem is a hierarchy of RAID levels arranged with faster, costlier and smaller tiers closer to the memory subsystem and slower, inexpensive and bigger tiers farther from the memory subsystem. The storage is adaptive in that data are migrated across the tiers depending on the access patterns, so that frequently accessed data are found in higher tiers and not so frequently accessed data are moved to the lower tiers. All this should be transparent to the application using the device. We first present the motivation and then discuss the issues in Adaptive hierarchical RAID.

3.2.1. *Motivation*

Typical nonscientific system disk access patterns show very high locality and the working set changes relatively slowly. This motivates one to consider a hierarchy of storage to exploit this locality. As we have seen above, RAID1 has best overall performance of all RAID levels. RAID5, on the other hand, has poor small update performance. This leads us to consider a hierarchy of RAID with a small RAID1 layer and a larger RAID5 layer.

3.2.2. *Issues in hierarchical RAID*

There are some issues that need to be addressed by any hierarchical storage system. We now present issues that are of significance to adaptive hierarchical RAID.

Transparency to the user: All the data migrations should be completely transparent to the application. This is required as the applications need not be changed to make use of the storage. This can be achieved by maintaining a *logical-physical* translation table. Also this translation table should be persistent across reboots.

Data organization at each level: The organization of data (or data layout) at each level in the hierarchy should be specified. For example, the prototype implementation has two levels of hierarchy, a declustered RAID1 level and a RAID5 level.

Data migrations: When should the data be migrated up (promoted) or down (demoted) the hierarchy should be specified. When no free space is available in the target level, some of the data in that layer should be demoted to lower level to make room for the data being promoted. The policy of deciding the victim data is called *victimization policy* and process of demoting it is called *victimization*. The hierarchical RAID storage should provide *reliable persistence semantics*, i.e. there should be no loss of data in the event of a system failure at any point during migrations.

3.3. *Comparison with previous work*

Previous work on adaptive hierarchical RAID such as HP AutoRAID[2] has explored one part of the design space, namely, design of configurable storage at the SCSI level with no interaction with higher-level layers like volume manager. This work explores a different design point, namely, one that is centered at the volume manager layer. This is important also for the reason that with fiber channel disks and SCSI-3, storage area networks (SAN) no longer need a conventional controller but a modified version of a controller that is more close to a volume manager.

This work shows that a hierarchy of RAID1 and RAID5 can be employed to exploit the locality of reference in the access patterns of typical systems. A prototype of adaptive hierarchical RAID has been implemented as a layered device driver on Sun Solaris 2.5.1. The work explains the advantages of such an approach over hardware implementation and also the limitations of a software implementation. An accidental side effect of the design is that implementing transaction semantics is fairly easy as compared to RAID5. The experimental results confirm the claim.

References

1. CHEN, P M , LEE, E K , GIBSON, G A , KALZ, R. H. AND PATTERSON, D A — High-performance, reliable secondary storage, *ACM Computing Surv* , 1994, **26**, 145–185

2 WILKES, J., GOLDING, R , STACLIN, C AND SULLIVAN, C — The HP autoraid hierarchial storage system, *Proc ACMC Symp. on Operating Systems Principles*, Dec 1995, pp. 96–108.